Félig kompozicionális szerkezetek automatikus felismerése doménadaptációs technikák segítségével a Szeged Korpuszon

Nagy T. István[1], Vincze Veronika[2], Zsibrita János[1]

[1]Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2., e-mail:{nistvan,zsibrita}@inf.u-szeged.hu
[2]Magyar Tudományos Akadémia, Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103., e-mail:vinczev@inf.u-szeged.hu

Kivonat Jelen tanulmányunkban bemutatjuk megközelítésünket, mely félig kompozicionális szerkezeteket képes automatikusan azonosítani magyar nyelvű szövegekben. Első lépésben a lehetséges jelölteket találjuk meg a szövegben, majd egy gazdag jellemzőkészleten alapuló bináris osztályozó segítségével azonosítjuk az egyes félig kompozicionális szerkezeteket. Módszerünket a Szeged Korpusz öt különböző doménjén is megvizsgáljuk, valamint két hasonlósági gráf segítségével azonosítjuk az egymáshoz közel álló részkorpuszokat. A különböző doméneken való vizsgálódások során egy egyszerű doménadaptációs módszert is bemutatunk.

1. Bevezetés

Az olyan főnévből és igéből álló többszavas kifejezéseket, ahol a szemantikai fej a főnév, míg az ige csupán a szerkezet igeiségéért felel, félig kompozicionális szerkezeteknek (FX-ek) nevezzük. Mivel ezen szerkezetek jelentése nem teljesen kompozicionális, ezért azok elemeinek egyenkénti lefordítása nem (vagy csak nagyon ritkán) eredményezi a szerkezet idegen nyelvű megfelelőjét. Az FX-ek automatikus azonosítását továbbá jelentősen megnehezíti, hogy e típusú összetett szerkezetek szintaktikailag hasonló felépítéssel bírnak (*választ kap*), mint más produktív (kompozicionális) szerkezetek (*pulóvert kap*), illetve idiómák (*vérszemet kap*) [1]. Az angol vonzatos igékhez (phrasal verbs) hasonlóan, célszerű az FX-eket is egyetlen komplex egységként kezelni azok nyelvi elemzésekor, hiszen a szerkezet szintaktikai és szemantikai feje nem azonos [2].

Jelen előadásban gépi tanulási megközelítésen alapuló módszerünket ismertetjük, mely magyar nyelven képes a félig kompozicionális szerkezetek automatikus azonosítására folyó szövegben. Továbbá megvizsgáljuk az általunk meghatározott szintaktikai elemzésen alapuló FX-jelöltkiválasztó módszer hatékonyságát. Gépi tanuló megközelítésünk az általunk leírt gazdag jellemzőtéren alapszik, mely egyaránt alkalmaz felszíni jellemzőket, szófaji információkat, funkcióigelistát, valamint szintaktikai és szemantikai információkat.

Módszerünk hatékonyságát a Szeged Korpusz [3] öt különböző doménén (jogi szövegek, fogalmazások, szépirodalmi szövegek, üzleti rövidhírek, újságcikkek) vizsgáltuk meg, melyeken az egyes FX-előfordulások manuálisan annotálva vannak. Mivel úgy találtuk, hogy különböző típusú szövegek különböző típusú félig kompozicionális szerkezeteket tartalmazhatnak, továbbá az FX-ek gyakorisága is eltérhet az egyes doméneken, ezért annak érdekében, hogy ezen különbségeket áthidaljuk, különös figyelmet fordítottunk az egyes korpuszokon tanult modellek hordozhatóságára, melyet egyszerű doménadaptációs technika segítségével valósítottunk meg. Az egyes szövegtípusok közti különbségek bemutatására a különböző doméneken előforduló félig kompozicionális szerkezetek gyakoriságából számított Kendall-együtthatót alkalmaztuk. Ezen domének közti eltéréseket a gépi tanuló algoritmusok által épített modellek által elért eredmények is alátámasztják.

2. Kapcsolódó munkák

Több megközelítést is implementáltak már félig kompozicionális szerkezetek automatikus azonosítására, valamint főnév + ige szerkezetek különböző osztályokba sorolására. Ezek közül a legtöbben alapvetően ige-tárgy párokra koncentráltak, amikor FX-et próbáltak azonosítani. A nem angol nyelvű kutatások során gyakran ige-prepozíció-főnév szerkezeteket vizsgáltak, mint például Van de Cruys és Moirón [4], akik holland nyelvű FX-ek azonosítása során alapvetően szemantikai jellemzőket felhasználó megközelítést alkalmaztak.

Számos megközelítés, mint például Stevenson és társai [5], valamint Van de Cruys és Moirón [4] alapvetően statisztikai jellemzőkre támaszkodva próbált meg automatikusan FX-et azonosítani. Ahogy Vincze [2] is rámutat, egy adott korpuszban az FX-ek nagy többsége igen ritkán fordul elő egy adott korpuszon. A vizsgált nagyméretű szövegeken az FX-ek 87%-a fordul elő kevesebb mint háromszor, ennél fogva igen nehéz pusztán statisztikai jellemzők alapján azonosítani őket.

Diab és Bhutada [6], valamint Nagy T. és társai [7] jellemzően (sekély) nyelvi információkra támaszkodó szabályalapú rendszereket alkalmaztak FX-ek azonosítására. Vincze és társai [8] szabályalapú rendszerüket mind magyar, mind angol nyelven alkalmazták többek közt a SzegedParallelFX párhuzamos korpuszon.

Statisztikai és nyelvi információkat egyaránt felhasználó rendszert építettek többek közt Tan és társai [9], valamint Tu és Roth [10]. Mindkét megközelítés ige + főnév párokat osztályoz aszerint, hogy félig kompozionális szerkezet-e vagy sem. Tu és Roth mind környezeti, mind statisztikai jellemzőket felhasználva tanított egy támasztóvektorgép-modellt a pozitív és negatív példák számában kiegyensúlyozott adathalmazon. Tanulmányuk szerint a többértelmű példákon a lokális jellemzőket használva érhetünk el jobb eredményeket. A Tan és társai által alkalmazott gépi tanuló alkalmazás statisztikai, valamint nyelvi információkat kombinálva véletlen erdő módszerét alkalmazva osztályozta a lehetséges FX-jelölteket.

Az általunk megvalósított megközelítés szintaktikai jellemzők alapján automatikusan kinyert főnév + ige párokat osztályoz gazdag jellemzőtérre támaszkodó gépi tanuló módszer alapján.

3. A félig kompozicionális szerkezetek automatikus azonosítása

Jelen munkában elsődleges célunk minden félig kompozicionális szerkezet automatikus azonosítása magyar nyelvű folyó szövegekben.

Mivel a különböző típusú szövegek merőben eltérő félig kompozicionális szerkezeteket tartalmazhatnak, valamint a különböző szövegekben más-más arányban fordulhatnak elő ezen szerkezetek, ezért fontosnak találtuk megvizsgálni az egyes doménen tanult modellek hordozhatóságát. Ezért módszereink kiértékelésére a Szeged Korpuszt használtuk, melyen öt különböző típusú szövegben vannak a félig kompozicionális szerkezetek manuálisan annotálva. Habár a korpuszban az FX-ek melléknévi igenévi és főnévi alakjai is jelölve vannak, mi alapvetően csak az igei alakok felismerésére fókuszáltunk. A Szeged Korpusz adatai az 1. táblázatban találhatók.

1. táblázat. A Szeged Korpusz adatai.

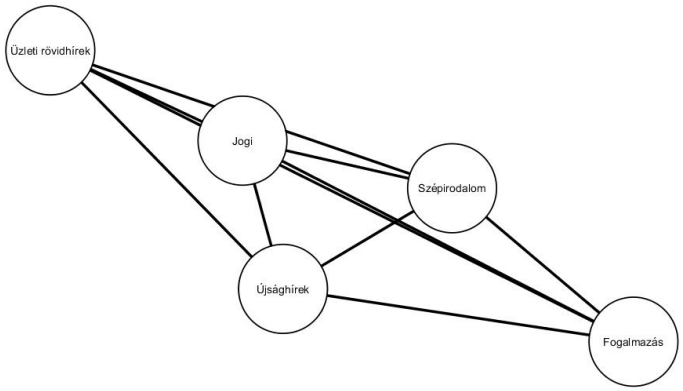
Korpusz	**Mondatok száma**	**Tokenek száma**	**FX**
Fogalmazás	23136	314787	677
Jogi	7058	188899	698
Szépirodalom	17358	219784	634
Üzleti rövidhírek	8956	213936	582
Újsághírek	8848	191156	484
Összesen	65356	1128562	3075

Mivel az alkalmazott megközelítésünk nagymértékben támaszkodik a szintaktikai jellemzőkre, ezért a Szeged Korpusznak csak azon részét használtuk fel, melyre a `magyarlanc 2.0` [11] szintaktikai elemzést tudott adni. Így végül öt különböző doménen 65356 mondaton 3075 FX-et vizsgáltunk. Az egyes részkorpuszokon tízszeres keresztvalidációval tanított és predikált modellek szófaji és függőségi elemzését használtuk. Mivel az etalon szófaji és függőségi elemzések egyaránt elérhetőek a Szeged Korpuszon, ezért lehetőségünk nyílt megvizsgálni, milyen hatással vannak a `magyarlanc 2.0` által nyújtott automatikus nyelvi elemzések megközelítésünk eredményességére. A különböző domének összehasonlítására kiszámoltuk az egyes részkorpuszokon a 15 leggyakrabban előforduló félig kompozicionális szerkezet Kendall-konkordancia értékeit, melyek a 2. táblázatban láthatóak.

A Kendall-együttthatók értékei alapján az egyes részkorpuszok hasonlóságát a 1. ábrán látható doménhasonlósági gráf segítségével ábrázoltuk, ahol az FX-

2. táblázat. Részkorpuszok Kendall-konkordancia értékei a 15 leggyakrabban előforduló félig kompozicionális szerkezet alapján.

-	**Fogalmazás**	**Jogi**	**Szépirodalom**	**Üzleti rövidhírek**	**Újsághírek** -
Fogalmazás	1	0,1825	0,5883	0,064	0,2498
Jogi	0,1825	1	0,2849	0,5068	0,3922
Szépirodalom	0,5883	0,2849	1	0,2422	0,2417
Üzleti rövidhírek	0,064	0,5069	0,2422	1	0,2409
Újsághírek	0,2498	0,3922	0,2417	0,2409	1

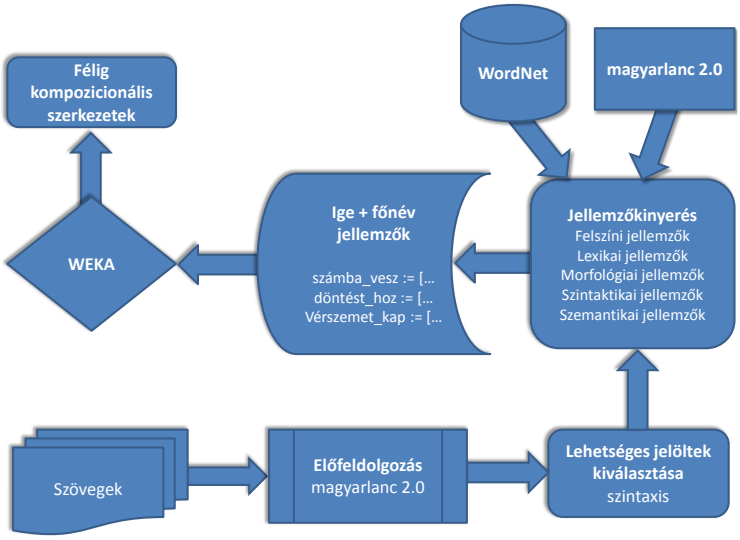


1. ábra. Doménhasonlósági gráf Kendall-együttható alapján.

ek szempontjából hasonló típusú szövegek közelebb, míg a kevésbé hasonlóak távolabb helyezkednek el egymástól.

3.1. Gépi tanuló megközelítés félig kompozicionális szerkezetek automatikus azonosítására

A félig kompozicionális szerkezetek automatikus azonosítására egy gépi tanuló megközelítést implementáltunk. Ehhez első lépésben minden mondatot elemzünk, és a lehetséges félig kompozicionális szerkezeteket szintaxisalapú jelöltkiválasztó megközelítés segítségével automatikusan kinyerjük. A második lépésben egy gazdag jellemzőkészleten alapuló bináris osztályozó segítségével döntünk, hogy egy adott potenciális szerkezet valóban félig kompozicionális szerkezet-e vagy sem. A 2. ábra mutatja be a teljes rendszer működését.



2. ábra. Rendszerábra.

3.2. Automatikus jelöltkinyerés

Azáltal, hogy az egyes félig kompozicionális szerkezetek a Szeged Korpusz részkorpuszain manuálisan annotálva vannak, lehetőségünk nyílt megvizsgálni ezen szerkezetek szintaktikai kapcsolatait folyó szövegekben. Ezen vizsgálataink alapján a lehetséges félig kompozicionális szerkezetekre úgy tekintettünk, mint olyan ige-főnév párok, melyek közt `subj`, `obj`, vagy `obl` (alany, tárgy vagy egyéb argumentum) szintaktikai kapcsolat van. Ahogy a 3. táblázatban látható, ezzel a jelöltkinyerő megközelítéssel képesek vagyunk a félig kompozicionális szerkezetek 92,07%-át automatikusan azonosítani.

3. táblázat. Az egyes részkorpuszokon előforduló félig kompozicionális szerkezetek szintaktikai kapcsolatai.

Korpusz	**OBJ**	**OBL**	**SUBJ**	**Összesen**	**Etalon**	**Fedés %**
Fogalmazás	401	171	45	617	677	91,14%
Jogi	394	150	97	641	698	91,83%
Szépirodalom	296	257	27	580	634	91,48%
Üzleti rövidhírek	339	176	19	534	582	91,75%
Újsághírek	307	130	22	459	484	94,83%
Összesen	1737	884	210	2831	3075	92,07%

3.3. Gépi tanuló alapú automatikus jelöltosztályozás

A következőkben bemutatjuk gépi tanuló alapú megközelítésünket, amelyet a lehetséges félig kompozionális szerkezetek automatikus osztályozására implementáltunk, és amely a következő osztályokba sorolható gazdag jellemzőkészleten alapszik: felszíni, lexikai, morfológiai, szintaktikai és szemantikai.

- Felszíni jellemzők: a **végződés** jellemző azt vizsgálja, hogy a szerkezet főnévi tagja bizonyos bi- vagy trigramra végződik-e. Ezen jellemző alapja, hogy az FX-ek főnévi komponense igen gyakran egy igéből képzett főnév. A szerkezetet alkotó **tokenek száma** szintén jellemzőként lett felhasználva.
- Lexikai jellemzők: A **leggyakoribb ige** jellemző az FX-ek azon tulajdonságát használja fel, hogy általában a leggyakoribb igék szerepelnek funkcióigeként (például *ad*, *vesz*, *hoz* stb.). Ezért az FX-jelöltek igei komponensének lemmáját vizsgáltuk, hogy az megegyezik-e az előre megadott leggyakoribb igék egyikével. A SzegedParalellFX korpuszban manuális annotált FX-ből gyűjtött, lemmatizált **FX lista** is felhasználásra került mint bináris jellemző, amely akkor kapott igaz értéket, ha az adott potenciális FX szerepelt a listában.
- Morfológiai jellemzők: mivel a magyar nyelv igen gazdag morfológiával rendelkezik, ezért számos morfológiaalapú jellemzőt definiáltunk. A **POS** módszernél FX-ekre jellemző szófaji mintákat definiáltunk, és amennyiben az FX-jelöltre illeszkedett egy minta, a jellemző igaz értéket kapott. További jellemzőként definiáltuk a funkcióigék **MSD-kódját** felhasználva az ige módját (`Mood`), valamint a főnévi komponens típusát (`SubPos`), esetét (`Cas`), a birtokos számát (`NumP`), a birtokos személyét (`PerP`), valamint a birtok(olt) számát (`NumPd`). A **szótő** jellemző alapvetően a főnévi komponens szótövét vizsgálja. Ez a jellemző az FX-ek azon már említett tulajdonságát kívánja kihasználni, hogy a félig kompozicionális szerkezetek főnévi tagja igen gyakran egy igéből származik, ezért azt vizsgáltuk, hogy a főnév tag szótövének van-e igei elemzése.
- Szintaktikai jellemzők: potenciális FX-ek kiválasztásánál alapvetően **szintaktikai információkra** támaszkodtunk. Ugyanakkor jellemzőként definiáltuk, hogy a három szintaktikai osztály (alany, tárgy vagy egyéb) melyike áll fenn az aktuális FX-jelölt esetében.
- Szemantikai jellemzők: ebben az esetben is az FX azon tulajdonságát használtuk fel, hogy a főnévi tag igen gyakran egy igéből származik. Ezért a Magyar WordNet-et [12] felhasználva `tevékenység` vagy `esemény` **szemantikai jelentést** keresünk a főnévi tag felsőbb szintű hipernimái közt.

Mivel a fentebb ismertetett jellemzők nagy része bináris attribútum, ezért a WEKA [13] csomagban elérhető, a C4.5 [14] döntési fa algoritmust implementáló J48 tanuló algoritmust alkalmaztuk. Rendszerünket minden részkorpuszon mondatszintű tízszeres keresztvalidációval értékeltük ki. A kiértékelés során a pontosság, fedés és F-mérték metrikákat használtunk. Ahogy a 3. táblázatban is látható, a potenciális FX-jelölt kiválasztó megközelítésünk az egyes korpuszokban manuálisan annotált FX-k 92,07%-át fedi csak le, ezért a gépi tanuló megközelítések fedés eredményeit korrigálnunk kellett.

Az egyes részkorpuszok összehasonlítására egyszerű, domének közötti keresztméréseket alkalmaztunk, mely során a forráskorpuszon tanított modelleket értékeltük ki a célkorpuszokon. Tehát a tanítóhalmaz nem tartalmazott annotált mondatokat a célkorpuszról.

Amennyiben nagyobb számú etalon példa áll rendelkezésünkre más-más doménekről és csak korlátozott számú példával rendelkezünk a feladat szempontjából érdekes doménről, akkor doménadaptációs technikák segítségével javíthatjuk rendszerünk hatékonyságát. Vagyis hatékonyabb gépi tanuló modellt építhetünk, ha a nagyméretű forrásdomén tanítóhalmazt kiegészítjük a céldoménen elérhető kisebb etalon korpusszal.

A Szeged Korpusz öt különböző típusú részkorpuszának köszönhetően megvizsgálhattuk, hogy egyszerű doménadaptációs technikák segítségével hogyan növelhetjük rendszerünk teljesítményét. Egy nagyon egyszerű doménadaptációs megoldást alkalmaztunk: a tanítóhalmazt kiegészítettük 500 céldoménről véletlenszerűen kiválasztott mondattal, majd 500 mondatonként növeltük a céldoménről érkező mondatok számát egészen 3000-ig. A doménadaptáció kiértékelésére is mondatszintű tízszeres keresztvalidációt alkalmaztunk. Az eredmények összehasonlíthatósága érdekében a keresztvalidáció során ugyanazon teszthalmazokat alkalmaztuk a céldoménen, mint a doménen belüli kiértékelés során. Ugyanakkor figyelmet fordítottunk arra is, hogy a doménadaptációhoz véletlenszerűen kiválasztott mondatok egyike se szerepeljen az aktuális teszthalmazban.

Baseline megoldásnak szótárillesztési megközelítést vettünk. Minden részkorpusz esetében a gépi tanuló megközelítésben is alkalmazott, a SzegedParallelFX korpuszon manuálisan annotált FX-ekből létrehozott lista lemmatizált verzióját használtuk a szótárillesztés során. Amennyiben a lista egy eleme előfordult egy adott mondat lemmatizált verziójában, akkor azt FX-nek jelöltük. Az etalon, valamint predikált jellemzőket felhasznált gépi tanult modellek eredményei és a szótárillesztés eredményei a 4. táblázatban, míg a keresztmérések eredményei a 6. táblázatban találhatók.

4. Eredmények

A tízszeres keresztvalidációval kiértékelt eredmények alapján a jogi korpuszon értük el a legjobb eredményeket 68,35 F-mértékkel. Ugyanakkor a legnehezebb doménnek a fogalmazás (51,83 F-mérték) és az újsághírek (51,84 F-mérték) részkorpuszok bizonyultak. Az etalon és predikált jellemzőkön tanult gépi tanuló modellek közt a szépirodalmi korpuszon volt a legnagyobb, 1,5 pontos eltérés, míg az üzleti rövidhírek esetében csupán 0,23 pontos különbség mutatkozott. Az öt korpuszon átlagosan 0,69 ponttal bizonyultak jobbnak az etalon jellemzőket használó modellek a predikált jellemzőket használóknál. A szótárillesztés a fogalmazás doménen bizonyult a leghatékonyabbnak 32,91 pontos F-mértékkel, és szintén ezen a részkorpuszon mutatkozott a legkisebb eltérés a gépi tanuló modell és baseline megközelítés közt. Szemben a jogi doménnel, ahol a két megközelítés közt 41,76 pontos eltérés mutatkozott.

4. táblázat. Szótárillesztés, valamint a gépi tanult megközelítés eredményei a különböző doméneken, etalon és predikált jellemzőket felhasználva.

Korpusz	**Pontosság**	**Fedés**	**F-mérték**	**Különbség**
Fogalmazás				
etalon	53,05	50,66	51,83	-
predikált	54,18	48,74	51,32	-0,51
szótárillesztés	52,85	23,88	32,91	-18,92
Jogi				
etalon	68,65	68,05	68,35	-
predikált	68	66,91	67,45	-0,9
szótárillesztés	47,52	18,46	26,59	-41,76
Szépirodalom				
etalon	56,72	47,48	51,69	-
predikált	52,27	48,26	50,19	-1,5
szótárillesztés	68,81	23,71	35,26	-16,43
Üzleti rövidhírek				
etalon	65,04	57,9	61,26	-
predikált	62,51	59,62	61,03	-0,23
szótárillesztés	53,48	18,42	27,39	-33,87
Újsághírek				
etalon	49,56	54,34	51,84	-
predikált	51,17	51,86	51,51	-0,33
szótárillesztés	43,72	20,52	27,93	-23,91
Átlag				
etalon	49,56	54,34	56,99	-
predikált	57,63	55,08	56,3	-0,69
szótárillesztés	53,28	20,99	30,02	-26,97

5. táblázat. Az egyes jellemzőosztályok.

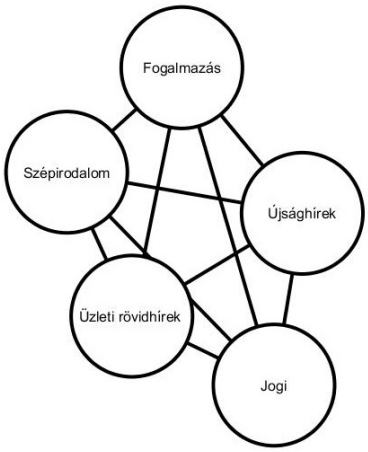
Jellemző	**Pontosság**	**Fedés**	**F-mérték**	**Eltérés**
Felszíni	53,73	56,19	54,93	-6,1
Lexikai	47,98	40,38	43,85	-17,18
Morfológiai	61,34	57,56	59,39	-1,64
Szintaktikai	61,35	59,11	60,21	-0,82
Szemantikai	63,4	56,76	59,9	-1,13
Összes	62,51	59,62	61,03	0

Hogy megvizsgálhassuk, az egyes jellemzők miként befolyásolják a gépi tanuló rendszer eredményeit, az üzleti rövidhír részkorpuszon porlasztásos mérést végeztünk, melynek eredményei a 5. táblázatban láthatók. Ekkor a teljes jellemzőtérből elhagytuk az egyes jellemzőcsoportokat, majd a maradék jellemzőkre támaszkodva tanítottunk. Az eredmények alapján a leghasznosabbnak a lexikai, valamint a felszíni jellemzők bizonyultak. A lexikai jellemzők közül elsősorban a funkcióige-lista bizonyult a leghatékonyabb jellemzőnek.

A keresztmérések alapján, a fogalmazás korpuszon a szépirodalmi doménen tanított modell teljesített a legjobban 43,29 pontos F-mértéket elérve. Ugyan 11,96 ponttal kisebb F-mértéket tudott elérni az üzleti rövidhíreken tanult modell a jogi részkorpuszon a céldoménhez képest, ám így is ez a modell volt a leghatékonyabb a többi közül. A szépirodalmi doménen a fogalmazás korpuszon tanult megközelítése bizonyult a legjobbnak 49,84 pontos F-mértékkel. Üzleti rövidhírek esetében a legjobb eredményt az újsághíreken tanított gépi tanulási modell érte el 55,75 pontos F-mértékkel. 50,42 pontos F-mértékkel az üzleti rövidhíreken tanított, ám az újsághíreken predikált modell bizonyult a legjobbnak.

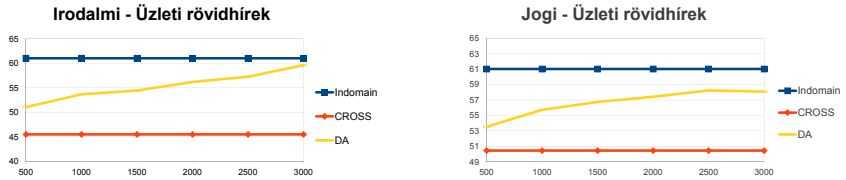
6. táblázat. Keresztmérések eredményei az egyes részkorpuszokon.

Korpusz	**Pontosság**	**Fedés**	**F-mérték**	**Eltérés**
Fogalmazás	54,18	48,74	51,32	-
Jogi	20,08	39,44	26,61	-24,71
Szépirodalom	37,62	50,96	**43,29**	-8,03
Üzleti rövidhírek	37,31	36,93	37,12	-14,02
Újsághírek	37,62	29,39	33	-18,32
Jogi	68	66,91	67,45	-
Szépirodalom	52,98	47,13	49,89	-17,56
Fogalmazás	55,21	40,26	46,56	-20,89
Üzleti rövidhírek	64,22	48,85	**55,49**	-11,96
Újsághírek	69,18	42,12	52,36	-15,09
Szépirodalom	52,27	48,26	50,19	-
Jogi	27,92	32,81	30,17	-20,02
Fogalmazás	60,75	42,19	**49,84**	-0,35
Üzleti rövidhírek	51,04	38,64	43,99	-6,2
Újsághírek	42,04	20,82	27,85	-22,34
Üzleti rövidhírek	62,51	59,62	61,03	-
Jogi	43,89	59,28	50,44	-10,59
Szépirodalom	40,85	51,37	45,51	-15,52
Fogalmazás	48,22	34,88	40,48	-20,55
Újsághírek	60	52,06	**55,75**	-5,28
Újsághírek	51,17	51,86	51,51	-
Jogi	30,76	61,78	41,07	-10,44
Szépirodalom	34,8	55,58	42,8	-8,71
Fogalmazás	40,64	41,74	41,18	-10,33
Üzleti rövidhírek	46,29	55,37	**50,42**	-1,09



3. ábra. Doménhasonlósági gráf keresztmérések eredményei alapján.

A keresztmérések eredményei alapján az egyes domének közti hasonlóságokat a 3. ábrán látható irányítatlan, súlyozott gráf segítségével jelenítettük meg. A gráf súlyait az adott domén tízszeres keresztvalidációval mért eredményei, valamint a keresztmérések különbségei adták.



4. ábra. Doménadaptációs eredmények üzleti rövidhírek doménen, irodalmi és jogi részkorpuszon tanítva.

A doménadaptációs mérések eredményei a 4. ábrán látható. A két kép bemutatja, hogy az adaptációhoz használt mondatok számának változásával hogyan módosul az adott doménen a rendszer által elért F-mérték.

Mind a két esetben jól látszik, hogy az adaptációhoz a céldoménről felhasznált mondatok számával folyamatosan növekednek a céldoménen elért eredmények. Az irodalmi részkorpuszt forrásdoménként használva, a doménadaptáció segítségével a céldoménen tízszeres keresztvalidációval elérhető eredményét közelítettük

meg. A doménadaptáció határozottan képes volt javítani a jogi részkorpusz forrásdoménről történő keresztmérés eredményéhez képest.

5. Az eredmények értékelése, összegzés

Jelen munkánkban bemutattuk gazdag jellemzőtéren alapuló gépi tanuló megközelítésünket, mely automatikusan képes magyar nyelvű szövegekben félig kompozicionális szerkezeteket azonosítani. A problémát két lépésből álló megközelítéssel oldottuk meg: az első lépésben a folyó szöveg mondataiból a potenciális FX-jelölteket nyertük ki automatikusan, egy alapvetően szintaxisra támaszkodó jelöltkiválasztó megközelítéssel. Módszerünk igen hatékonynak bizonyult, mivel a manuálisan annotált FX-ek 92%-át sikerült lefedje. A kinyert példák közül automatikusan azonosítottuk az egyes FX-eket egy gazdag jellemzőtéren alapuló bináris osztályozó segítségével. Módszerünket a Szeged Korpusz egyes doménjein értékeltük ki, azt vizsgálva, mely részkorpuszok hasonlítanak a leginkább egymásra, melyeken fordulnak elő hasonló FX-ek.

Az egyes domének közötti hasonlóságok kifejezésére két hasonlósági gráfot is megadtunk. Az első esetben az egyes részkorpuszokon előforduló FX-ek gyakoriságából számított Kendall-együtthatóval súlyoztuk a gráf egyes éleit, míg a másik esetben a keresztmérések eredményei alapján lettek a gráf élei súlyozva. Ezek alapján megállapítható, hogy a fogalmazás és a szépirodalom domének, valamint a újsághírek és üzleti hírek domének hasonlítanak egymásra a legjobban. A jogi szövegek pedig inkább az utóbbi két részkorpuszhoz hasonlítanak.

Rendszerünk hibaelemzése is alátámasztotta a porlasztásos mérés során is bemutatott eredményt, miszerint a leghatékonyabb jellemzőnek a funkcióige-lista bizonyult. Ugyanis a hibaelemzés során kiderült, hogy a helyesen predikált FX-ek igéinek több mint 80%-a szerepelt a funkcióige-listában, míg az álpozitív FX-ek igéinek kevesebb mint 10% volt megtalálható a listában. Az elemzés arra is enged következtetni, hogy rendszerünk alapvetően a rövidebb, kevesebb mint 3 tokenből álló FX-t azonosítja helyesen. Továbbá néhány álpozitív eredmény annotálási hibára, valamint helytelen szófajkódi elemzésre vezethető vissza.

Megközelítésünket különböző doméneken is kiértékeltük, az egyes részkorpuszokon elérhető eredményeket pedig egyszerű doménadaptációs technikákkal javítottuk. Eredményeink azt mutatják, hogy a magyar nyelvű FX-ek folyó szövegben való automatikus azonosítása igen kihívásokkal teli feladat, de az általunk bemutatott megközelítés erre a nehéz problémára nyújt egy lehetséges megoldást.

Köszönetnyilvánítás

Jelen kutatást a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt támogatta az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett.

Hivatkozások

1. Vincze, V.: Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus. In: Proceedings of LREC-2012, Istanbul, Turkey, ELRA (2012) 2381–2388
2. Vincze, V.: Semi-Compositional Noun + Verb Constructions: Theoretical Questions and Computational Linguistic Analyses. Doktori értekezés, Szeged, Szegedi Tudományegyetem (2011)
3. Alexin, Z., Gyimóthy, T., Hatvani, Cs., Tihanyi, L., Csirik, J., Bibok, K., Prószéky, G.: Manually annotated Hungarian corpus. In: Proceedings of EACL-2003 - Volume 2. EACL '03, Stroudsburg, PA, USA, ACL (2003) 53–56
4. Van de Cruys, T., Moirón, B.n.V.: Semantics-based multiword expression extraction. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions. MWE '07, Stroudsburg, PA, USA, ACL (2007) 25–32
5. Stevenson, S., Fazly, A., North, R.: Statistical measures of the semi-productivity of light verb constructions. In: Proceedings of the Workshop on Multiword Expressions: Integrating Processing. MWE '04, Stroudsburg, PA, USA, ACL (2004) 1–8
6. Diab, M.T., Bhutada, P.: Verb noun construction MWE token supervised classification. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications. MWE '09, Stroudsburg, PA, USA, ACL (2009) 17–22
7. Nagy T., I., Vincze, V., Berend, G.: Domain-Dependent Identification of Multiword Expressions. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., eds.: RANLP, RANLP 2011 Organising Committee (2011) 622–627
8. Vincze, V., Nagy T., I., Zsibrita, J.: Félig kompozicionális szerkezetek automatikus azonosítása magyar és angol nyelven. In Tanács, A., Vincze, V., eds.: VIII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2011) 59–70
9. Tan, Y.F., Kan, M.Y., Cui, H.: Extending corpus-based identification of light verb constructions using a supervised learning framework. In: Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts, Trento, Italy, ACL (2006) 49–56
10. Tu, Y., Roth, D.: Learning English Light Verb Constructions: Contextual or Statistical. In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, Portland, Oregon, USA, ACL (2011) 31–39
11. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In Tanács, A., Vincze, V., eds.: MSzNy 2013 – IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2013) 368–374
12. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P., eds.: Proceedings of the Fourth Global WordNet Conference (GWC 2008), Szeged, University of Szeged (2008) 311–320
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1) (2009) 10–18
14. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)